Reducing Sensitivity on Speaker Names for Text Generation from Dialogues

Qi Jia[1], **Haifeng Tang**[2], **Kenny Q. Zhu**[3*]
[1,3]Shanghai Jiao Tong University, Shanghai, China
[2]China Merchants Bank Credit Card Center, Shanghai, China
[1]`Jia_qi@sjtu.edu.cn`, [2]`thfeng@cmbchina.com`, [3]`kzhu@cs.sjtu.edu.cn`

Abstract

Changing speaker names consistently throughout a dialogue should not affect its meaning and corresponding outputs for text generation from dialogues. However, pre-trained language models, serving as the backbone for dialogue-processing tasks, have shown to be sensitive to nuances. This may result in unfairness in real-world applications. No comprehensive analysis of this problem has been done in the past. In this work, we propose to quantitatively measure a model's sensitivity on speaker names, and comprehensively evaluate a number of known methods for reducing speaker name sensitivity, including a novel approach of our own. Extensive experiments on multiple datasets provide a benchmark for this problem and show the favorable performance of our approach in sensitivity reduction and quality of generation.

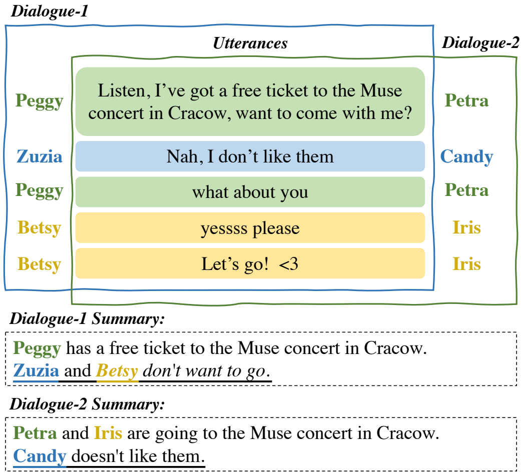

Dialogue-1
Utterances
Dialogue-2
Peggy
Listen, I've got a free ticket to the Muse concert in Cracow, want to come with me?
Petra
Zuzia
Nah, I don't like them
Candy
Peggy
what about you
Petra
Betsy
yessss please
Iris
Betsy
Let's go! <3
Iris
Dialogue-1 Summary:
Peggy has a free ticket to the Muse concert in Cracow.
Zuzia and Betsy don't want to go.
Dialogue-2 Summary:
Petra and Iris are going to the Muse concert in Cracow.
Candy doesn't like them.

Figure 1: Two instances of an example from the SAMSum dataset, each with a different set of names. Two different summaries are generated by BART. Different colors indicate different speakers. divergent contents are underlined and *incorrect contents* are italicized.

1 Introduction

The safety and fairness issue of generations from dialogue models is a crucial concern in real applications. Previous work focuses on response generation from open-ended dialogue systems (Xu et al., 2020; Henderson et al., 2018), such as offensive contents (Baheti et al., 2021), gender bias (Liu et al., 2020; Dinan et al., 2020) and other discriminated behavior (Sheng et al., 2021; Smith and Williams, 2021). For other text generation tasks where the whole dialogue is provided and the output shouldn't go beyond the dialogue, such as dialogue summarization (Gliwa et al., 2019) and dialogue reading comprehension (Li et al., 2020), the fairness issue is still unexplored.

In these tasks, the input dialogues are self-contained, and the names of the speakers do not carry any connotation from outside of the dialogue. Therefore, changing the speaker names consistently in a dialogue should not affect the meanings of the dialogue and the desired outputs. This contrasts with response generation, where the dialogue is in progress and the output is expected to be different in styles or contents for various speakers. Taking dialogue summarization (Gliwa et al., 2019; Chen et al., 2021) as an example for text generation from dialogues, it focuses on generating concise "who-did-what" summaries in the third person. In Fig. 1, the two dialogues are identical except for the speaker names. The two summaries are expected to be the same modulo the speaker names.

Unfortunately, models nowadays, following the pretrain-finetune paradigm, are sensitive to trivial changes, which has been verified in other tasks. In relation extraction, spurious correlations between entity mentions and relations lead to entity bias (Zhang et al., 2018, 2017; Wang et al., 2022b). Other similar work includes the analysis of robustness by entity renaming for machine reading comprehension models on narrative texts (Yan et al., 2022) and name biases in machine transla-

[*] The corresponding author.

tion with inflected languages (Wang et al., 2022a), like German. Besides, Shwartz et al. (2020) claims that pre-trained language models do not treat given names as interchangeable or anonymous, showing unfairness in reading comprehension.

Obviously, dialogue understanding models are sensitive to speaker names according to Fig. 1 as well. The model tends to generate different information given different speaker names, such as "don't want to go" and "doesn't like them". Incorrect content, "... Betsy don't want to go", is generated with the first group of speakers, while not with the other group. According to our pilot experiment with the vanilla BART fine-tuned on SAMSum, around 74.00% of generations are changed by switching speaker names and 69.82% among them are due to distinct contents. Such uneven performances create unfairness among different speakers, especially in the aspect of information allocation. The model may also catch latent properties in names (Romanov et al., 2019) and lead to discrimination, raising the importance of research on the sensitivity on speaker names.

Previous work has also mentioned this problem. Different data pre-processing approaches are adopted during the construction of datasets to avoid using speaker names, such as "A" or "B" in Li et al. (2017). Khalifa et al. (2021) replace speaker names with more common and frequent names that the model may have seen during pre-training. Data augmentation by changing speaker names is adopted by Liu and Chen (2021). However, all of them only attempted to attack this problem subjectively, without quantitive analysis and fair comparisons.

In this work, we systematically analyze speaker name sensitivity in text generation from dialogues. We define the speaker name sensitivity and divide the approaches into offline and online ones. Then, we propose two novel insensitivity losses, helping to reduce attention and hidden state distances of the same dialogue with different speaker names for transformer-based models during fine-tuning. These losses can be used in both kinds of approaches. Results on several tasks show that our losses reduce the sensitivity and get better generations. In summary, our contributions are:

- We are the first to investigate the speaker name sensitivity in text generation from dialogues (Sec. 2.1) with all of the codes and results open-sourced at `https://github.com/JiaQiSJTU/SpeakerNameSensitivity`.
- We introduce two novel insensitivity losses as auxiliary training objectives for reducing sensitivity during fine-tuning (Sec. 3).
- Experiments on different tasks provide a benchmark with comprehensive analysis on speaker name sensitivity, and show state-of-the-art performances of our approach (Sec. 5).

2 Background

2.1 Speaker Name Sensitivity

speaker name sensitivity is the differences in the generations by a model, given the identical dialogues except for different speaker names. We define it as follows.

Let d denote the input dialogue. c denotes other input content, which can be empty for tasks like dialogue summarization, or a piece of text such as a question for reading comprehension. p refers to the set of speakers names in d. f is a one-to-one mapping which maps p into a set of names p' from a name pool $\mathcal{P}$ consisting of a set of candidate names to be substituted into the samples. The names p' are sampled under the uniform distribution without the loss of generality. *The speaker name sensitivity SS of a generation model $\mathcal{M}(\cdot)$ on this sample* is:

$$\begin{aligned} SS(\mathcal{M}|d,c) = \delta(\{\mathcal{M}(Rep(d,c|f)) \\ |\forall f: p \rightarrow p', p' \subseteq \mathcal{P}\}) \end{aligned} \tag{1}$$

where $Rep(\cdot)$ replaces names in the sample given f, i.e., from p to p'. $\delta(\cdot)$ quantifies the differences among generations.

Then, *the sensitivity SS of a model $\mathcal{M}(\cdot)$* is the expectation $\mathbb{E}$ of over all samples from the real-world distribution D:

$$SS(\mathcal{M}) = \mathbb{E}_{(d,c)\sim D}[SS(\mathcal{M}|d,c)] \tag{2}$$

In practice, a dialogue dataset is regarded as a sampling from D for evaluations. Each sample in the dataset is provided with a reference output o for supervised training. We use D_{tr}, D_{va} and D_{te} to refer to training, validation and test sets. See detailed implementation and metrics in Sec. 4.1.

2.2 Existing Approaches

We investigate existing approaches that target on reducing the sensitivity and classify them into offline ones and online ones, where the former chases to reduce the sensitivity by exploring better model parameters and the latter pursues insensitivity by unification or simplification of input data. Thus, data processing steps are required before inputting into the model and after the inference during the test time and speaker names in D_{tr}, D_{va} and D_{te} are all changed for online approaches. The model needs fine-tuning for both approaches.

Offline approaches include:

Embedding Layer(Emb): Similar to (Gu et al., 2020) and (He et al., 2021), an additional embedding layer can be adopted for representing whether the model should be sensitive to corresponding tokens. 2 embeddings are learned during fine-tuning.

Augmentation (Aug): Liu and Chen (2021) proposed to do data augmentation by exchanging speaker names in training samples with names from D_{tr}. They aim to reduce unexpected inductive bias caused by speaker names, which is similar to our goal. The model is fine-tuned with augmented training data while D_{va} and D_{te} remain unchanged.

Online approaches are:

ID: Some works (Cui et al., 2020; Li et al., 2017) replace speaker names with predefined IDs to avoid name bias. We use "Speaker[NUM]" similarly to Kim et al. (2019) and Chen et al. (2021), which is close to words seen during pre-training and fits different numbers of speakers. "[NUM]" is the index of a speaker's first occurrence.

Frequent (Fre): This refers to the approach proposed in Khalifa et al. (2021). They use 100 frequent male and 100 frequent female names online[1] as the pool P for sampling replacements. This approach can be combined with Aug into **FreAug**.

3 Proposed Approach

We focus on the widely-accepted encoder-decoder architecture for pre-trained generation models and design two auxiliary insensitivity losses to take full advantage of augmented data on top of Aug. Given the dialogue sample with different speaker names, a model outputs distinct generations due to its different internal behaviors. Therefore, penalizing unexpected internal differences should help the model behave consistently and reduce the sensitivity.

With this intuition, we propose the cross-attention loss and the decoder-hidden-state loss. The former corresponds to cross-attention distributions that help the decoder make a soft information selection among encoder hidden states at each step and should be similar with different speaker names. The latter is based on the final decoder hidden states which are expected to be the same under the default teacher-forcing training strategy except for the speaker name tokens. We didn't consider the encoder attentions since according to our pilot analysis of the vanilla BART, the cross attentions distance of the different predictions is around 1.5 times of the same ones. However, there are no differences in the encoder attentions. Other intermediate hidden states are excluded since they are all affected by different input embeddings of speaker names, except that the final decoder hidden states are sure to be the same.

3.1 Cross-attention Insensitivity Loss

We denote a model's input and output length, i.e., the number of tokens, as din and $dout$. During training, the cross attentions calculated for each output token are collected as $CA \in R^{N\times dout\times din}$. N is the number of heads for the multi-head attention mechanism, determined by the configuration of pre-trained models. We apply average pooling over the dimension of $dout$, to get the overall attention over the input tokens $\overline{CA} \in R^{N\times din}$.

Given an original sample $\{d_i, c_i, o_i\}$, we construct $K-1$ augmented samples by replacing speaker names. The averaged attentions for all samples are $\{\overline{CA_k}\}_{k=1}^{K}$. Since it is a default that each sample should go through the tokenizer before inputting to the model, $\{din_k\}_{k=1}^{K}$ are not guaranteed to be identical in two cases. First, names may be tokenized into different token counts. For example, "John" and "Robinson" are tokenized into {"John"} and {"Rob", "inson"} by BART tokenizer. Replacing "John" with "Robinson" in d_i will increase the sequence length. Second, long inputs may be truncated at different tokens. So, we consider two corresponding functions for unification:

- $\text{Sum}(\cdot)$ sums up the attention values of tokens belonging to an occurrence of a speaker name.
- $\text{Pad}(\cdot)$ pads attentions into the same length din_u by concatenating zeros, which means that this part of contents is missing.

The unified $\{\overline{CA_k}\}_{k=1}^{K}$ is represented as $\{\widetilde{CA}_k\}_{k=1}^{K}$, where $\widetilde{CA}_k \in R^{N\times din_u}$.

Finally, the loss is calculated as:

$$\mathcal{L}_{ca} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{l=1, l\neq k}^{K} loss(\widetilde{CA}_k, \widetilde{CA}_l) \tag{3}$$

[1] https://www.ssa.gov/oact/babynames/decades/century.html

where $loss(\cdot)$ measures the distances between a pair of attentions.

3.2 Decoder-hidden-state Insensitivity Loss

Similarly, hidden states of the decoder's final output for all samples can be denoted as $\{DH_k\}_{k=1}^{K}$, where $DH_k \in R^{H\times dout_k}$ and H represents the hidden size. The lengths of them also vary due to the above two cases. We adopt two different functions:

- $\mathrm{Del}(\cdot)$ ignores the hidden states whose predicted tokens belong to a speaker name.
- $\mathrm{Trunc}(\cdot)$ truncates the redundant hidden states at the end without the paired ones.

Thus, the unified $\{DH_k\}_{k=1}^{K}$ is represented as $\{\widetilde{DH}_k\}_{k=1}^{K}$, where $\widetilde{DH}_k \in R^{H\times dout_u}$.

The loss is defined as:

$$\mathcal{L}_{dh} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{l=1,l\neq k}^{K} loss(\widetilde{DH}_k, \widetilde{DH}_l) \tag{4}$$

We adopted the mean square error for both losses.

3.3 Learning Objective

$\mathcal{L}_{ca}$ and $\mathcal{L}_{dh}$ are added to the vanilla generation loss $\mathcal{L}_{gen}$ with hyper-parameters α and β:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \alpha\mathcal{L}_{ca} + \beta\mathcal{L}_{dh} \tag{5}$$

The insensitivity losses are only auxiliary fine-tuning objectives, leaving the inference time unchanged. They can be added on top of both Aug and FreAug, denoted as **Ins** and **FreIns**.

4 Experimental Setup

We define the evaluation metrics for sensitivity, introduce multiple text generation tasks with dialogue data and present implementation details.

4.1 Evaluation Metrics for Sensitivity

We uniformly sample names from P, which is specified later, to realize f without the loss of generality and re-sample the name if it is not in p but in the conversation. We avoid changing names mentioned during the conversation in case they are grounded entities. Since it's impossible to enumerate all possible f, we choose to substitute names of samples in D_{te} for $T = 5$ times. It should be noted that varying names in test data is different from the augmentation approach. The additional test data is fixed once constructed for comparing approaches by quantitatively measuring the sensitivity.

We introduce three kinds of $\delta(\cdot)$ with task-specific evaluation metric $\mathrm{Score}(\cdot)$ and measure the speaker name sensitivity of a model similar to Prabhakaran et al. (2019)' work. **Pairwise Sensitivity(S-*)** is defined as:

$$E_{i=1}^{N^{te}} E_{t_1=1}^{T} E_{t_2=1,t_1\neq t_2}^{T} [1 - \mathrm{Score}(\hat{o}_i^{t_1}, \hat{o}_i^{t_2})] \tag{6}$$

$\hat{o}_i^t$ is the generation where replaced names are changed back for evaluation. N^{te} is the number of samples in D_{te}. $E(\cdot)$ is the mean operator.

Dialogue models are also expected to get the same scores with task-specific evaluation metrics compared with the reference o. So, we can also add o as the input of $\delta(\cdot)$ in Eq. 1 and define the following two metrics: **Score Range (R-*)** as

$$\begin{aligned} E_{i=1}^{N^{te}} [\, & \max(\{\mathrm{Score}(o_i, \hat{o}_i^t)|_{t=1}^{T}\}) \\ & - \min(\{\mathrm{Score}(o_i, \hat{o}_i^t)|_{t=1}^{T}\})] \end{aligned} \tag{7}$$

and **Score Deviation (D-*)** as

$$E_{i=1}^{N^{te}} [\mathrm{StdDev}(\{\mathrm{Score}(o_i, \hat{o}_i^t)|_{t=1}^{T}\})] \tag{8}$$

The sensitivity metrics here are the lower the better and are denoted by ↓ in the following sections.

4.2 Tasks and Datasets

We implement our experiments on the tasks below. The statistics are in Table 1 and we calculate the macro-average scores of samples for each metric.

Task	**Dialogue Summarization**	**Question Generation**	**Reading Comprehension**
Dataset	SAMSum	Molweni	Molweni
#Train	14,732	20,873	20,873
#Val	818	2,346	2,346
#Test	819	2,560	2,560
Output Length	23.44±12.72	7.05±2.02	4.01±2.93

Table 1: A summary of tasks. #Train, #Val and #Test refer to the number of samples in the datasets. Output length are statistics(avg±std) for the word counts.

Dialogue Summarization outputs fluent and concise summaries covering the salient information in dialogues. We experiment with the SAMSum dataset (Gliwa et al., 2019) consisting of around 16k open-domain dialogues among two or more interlocutors. Rouge-2 F1 (Lin, 2004) and BertScore F1 (Zhang et al., 2019)[2] are task-specific evaluation

[2] We adopted microsoft/deberta-xlarge-mnli recommended by https://github.com/Tiiiger/bert_score for BertScore.

metrics. We consider genders to be consistent when switching names following Khalifa et al. (2021).

Question Generation is to generate a question given an input dialogue and its corresponding answer span. We use Molweni dataset (Li et al., 2020) made up of around 10k task-oriented dialogues sampled from the Ubuntu Chat Corpus. Similar to the question generation work based on SQuAD1.1, we extract (dialogue, answer, question) tuples from the original Molweni dataset and ignore unanswerable questions. BLEU (Papineni et al., 2002) and Rouge-L F1 are used for evaluations.

Reading Comprehension generates an answer by inputting a dialogue with a question. We use the Molweni dataset (Li et al., 2020) and ignore unanswerable questions as well. Bleu and Rouge-L F1 are also used for evaluations.

4.3 Implementation Details

We use BART-large as our basic pre-trained model. We truncate inputs to the first 1024 tokens and the learning rate is $3e-5$ with weight decay equaling 0.01. The model is fine-tuned with batch size equaling 32 for 10 epochs. We evaluate the performance on D_{va} after each epoch with Rouge-2 F1 or Bleu. The checkpoint with the highest score on D_{va} is saved for testing. During the inference, we decode with no_repeat_ngram_size=3, length_penalty=1.0 and num_beams=4. We search α and β in {1, 10, 20} empirically and report results with the best validation performance. Specifically, α equals 1. β equals 1 for reading comprehension and 10 for the others. Our experiments are done on a single RTX 2080Ti with 11G GPU memory. Considering the GPU memory footprint, we set $K = 2$, which is the same for Aug and FreAug for fair comparisons.

We test online approaches with their corresponding test sets. For offline approaches, we focus on two sources of P. One is **in-distribution names** representing speaker names from the corresponding D_{tr}. The other is **all-possible names** with more than 117 thousand names[3], which can reflect the models' performances in complicated real scenarios. For approaches with sampling operations, we construct data with 3 different random seeds. Results are averaged over the number of runs.

5 Results

We show performances of approaches first, followed by ablation studies and human evaluations.

	Dialogue Summarization		Question Generation		Reading Comprehension	
Approach	**R2**	**BertS**	**Bleu**	**RL**	**Bleu**	**RL**
Vanilla	28.12	75.09	18.57	56.04	28.42	73.33
Emb	28.12	75.14	19.97	56.83	26.35	69.31
Aug	28.29	75.26	18.53	55.56	27.09	71.88
Ins⋆	**28.97**	**75.63**	**20.26**	**56.85**	**29.44**	**74.03**

Table 2: Performances(%) of offline approaches on the original test set. Vanilla refers to the baseline that simply fine-tuned the basic pre-trained model on the original dataset for different tasks. ⋆ marks our approach.

Then, we take a closer look at offline approaches, which show the inherent capability of models, with multi-faceted analysis. Hyper-parameter search and case studies are in Appendixes.

5.1 Performance of Offline Approaches

The performance on the original test sets is shown in Table 2. Emb only outperforms Vanilla on question generation and Aug only makes little improvements over Vanilla on dialogue summarization. Our approach Ins makes consistent improvements, performing best among offline approaches.

Results with sensitivity scores are in Table 3. Emb fails to generate more insensitive results, especially for question generation. Aug doesn't make promising improvements on outputs' quality over Vanilla, but it reduces the sensitiveness of models across different test sets and tasks. Ins leads to better results on randomly augmented training data with different random seeds, significantly outperforming Aug. In a word, Ins achieves the best performance among offline approaches.

By comparing the results in Table 3 horizontally, in-distribution names perform better than all-possible names on dialogue summarization, whereas results are opposite on the others. Speaker names in SAMSum are mostly real and popular names, while names in Molweni are online nicknames containing unknown words, such as "zykotick9". All-possible names contain a large proportion of real names, and a small proportion of names never seen during pre-training which can be regarded as nicknames. In this way, we can observe that the difficulty of modeling names for a model is "SAMSum in-distribution $<$ all-possible $<$ Molweni in-distribution". In other words, models perform better on more popular names, which is in accord with the success of Fre in Sec. 5.2.

5.2 Performance of Online Approaches

The results of online approaches are in Table 4.

[3] https://data.world/arunbabu/gender-by-names

	R2				BertScore			
Approach	-	S↓	R↓	D↓	-	S↓	R↓	D↓
In-distribution Names								
Vanilla	27.66	31.24	13.98	5.51	74.90	11.80	6.41	2.49
Emb	27.63	29.39	13.21	5.20	74.91	11.29	6.26	2.43
Aug	27.82	27.35	12.33	4.86	74.95	10.42	5.77	2.57
Ins⋆	**28.79**	**21.36**	**9.50**	**3.82**	**75.48**	**7.94**	**4.32**	**1.71**
All-possible Names								
Vanilla	27.19	33.10	14.64	5.72	74.83	12.26	6.66	2.60
Emb	27.22	31.38	13.59	5.30	74.89	12.03	6.63	2.55
Aug	27.50	28.17	12.56	4.97	74.96	10.56	5.76	2.25
Ins⋆	**28.44**	**25.37**	**11.58**	**4.62**	**75.38**	**9.38**	**5.22**	**2.05**

(a) Dialogue Summarization

	Bleu				RL			
Approach	-	S↓	R↓	D↓	-	S↓	R↓	D↓
In-distribution Names								
Vanilla	18.48	34.80	11.96	5.06	57.14	14.94	14.19	5.74
Emb	19.00	38.24	13.76	5.79	57.31	17.55	16.85	6.82
Aug	17.89	26.24	8.22	3.52	56.26	12.04	11.35	4.69
Ins⋆	**19.58**	**16.90**	**5.53**	**2.35**	**57.47**	**7.83**	**8.09**	**3.35**
All-possible Names								
Vanilla	18.56	29.64	10.04	4.26	57.38	12.98	11.88	4.90
Emb	18.70	35.52	12.55	5.27	57.28	16.05	15.26	6.20
Aug	17.81	23.09	7.15	3.06	56.08	10.66	9.64	4.03
Ins⋆	**19.57**	**14.65**	**4.41**	**1.90**	**57.49**	**6.96**	**6.58**	**2.78**

(b) Question Generation

	BLEU				RL			
Approach	-	S↓	R↓	D↓	-	S↓	R↓	D↓
In-distribution Names								
Vanilla	28.34	54.98	6.54	2.83	73.07	7.54	9.69	4.17
Emb	25.80	57.78	7.17	3.13	69.29	9.83	12.30	5.31
Aug	27.07	55.96	6.04	2.62	72.11	8.14	10.42	4.50
Ins⋆	**29.31**	**52.03**	**4.53**	**1.97**	**74.04**	**5.65**	**7.66**	**3.32**
All-possible Names								
Vanilla	28.56	53.94	5.39	2.34	73.60	6.39	8.21	3.53
Emb	25.99	56.22	5.11	2.21	69.59	7.29	8.60	3.69
Aug	27.12	54.72	5.15	2.23	72.23	6.39	8.29	3.58
Ins⋆	**29.34**	**51.38**	**3.66**	**1.59**	**74.35**	**4.62**	**6.15**	**2.64**

(c) Reading Comprehension

Table 3: Performances(%) of offline approaches. "-" is the original metric. S, D and R are shorted for the sensitivity metrics. Scores significantly better than all the baselines with p-value<0.05 are underlined.

	R2				BertScore			
Approach	-	S↓	R↓	D↓	-	S↓	R↓	D↓
ID	26.97	-	-	-	74.26	-	-	-
Fre	28.55	25.17	11.31	4.50	74.24	9.77	5.30	2.09
FreAug	27.86	25.03	11.09	4.39	75.02	9.58	5.12	2.02
FreIns⋆	**28.73**	**17.25**	**7.66**	**3.14**	**75.53**	**6.39**	**3.43**	**1.38**

(a) Dialogue Summarization

	BLEU				RL			
Approach	-	S↓	R↓	D↓	-	S↓	R↓	D↓
ID	19.21	-	-	-	56.49	-	-	-
Fre	18.96	18.44	5.51	2.35	57.10	8.35	7.23	3.04
FreAug	18.52	16.01	4.92	2.14	57.06	7.05	6.50	2.76
FreIns⋆	**19.71**	**10.09**	**3.12**	**1.35**	**57.29**	**4.48**	**4.19**	**1.80**

(b) Question Generation

	BLEU				RL			
Approach	-	S↓	R↓	D↓	-	S↓	R↓	D↓
ID	28.46	-	-	-	73.62	-	-	-
Fre	27.35	54.55	3.77	1.63	73.56	4.95	6.05	2.61
FreAug	27.92	52.67	3.28	1.42	73.67	4.24	5.63	2.43
FreIns⋆	**29.03**	**52.28**	**2.66**	**1.15**	**74.59**	**3.28**	**4.51**	**1.95**

(c) Reading Comprehension

Table 4: Performances(%) of online approaches.

All speaker names will be normalized into fixed code names in ID, so that the test set for ID is changeless for each sample and the sensitivity scores are actually 0.0. Unfortunately, its quality scores lag behind Ins and even drop dramatically on dialogue summarization. Thus, it's not recommended to be a necessary data pre-processing step.

Fre makes some improvements on R2 for dialogue summarization by comparing with the vanilla model, which is consistent with the results in (Khalifa et al., 2021), whereas the drops in BertScore were not mentioned in their work. The sensitivity scores are lower than those for offline approaches in Table 3. To better understand the gains of Fre, we further test the vanilla model with the same test sets replaced by frequent names. It achieves similar performance on Rouge-2 (28.18) and BertScore (75.13) with the vanilla model. The sensitivity score D-BertS is 2.24, which is lower than 2.49 of Vanilla in Table 3. It shows that the advantages of Fre not only come from using the group of frequent names that are easier for a model to understand, but also from doing fine-tuning with this group of names. FreAug doesn't improve the outputs' quality consistently, but reduces the sensitivity scores.

FreIns performs the most insensitively with better generation quality among online approaches.

5.3 Ablation Study

Ablation studies of our full approach Ins are in Table 5. Aug is regarded as an ablation representing the model trained without any auxiliary losses. Both insensitivity losses outperform Aug with using $\mathcal{L}_{dh}$ topping the rank on most metrics, showing that penalizing differences on the decoder hidden states has more direct effects on the outputs. Combining both losses induces more performance gains.

	Dialogue Summarization		Question Generation		Reading Comprehension	
Approach	**BertS**	**D-BertS↓**	**Bleu**	**D-Bleu↓**	**Bleu**	**D-Bleu↓**
Ins	**75.48**	**1.71**	19.48	**2.35**	**29.31**	**1.97**
-w/o $\mathcal{L}_{ca}$	75.43	1.85	**19.71**	2.47	29.03	2.19
-w/o $\mathcal{L}_{dh}$	74.89	2.27	18.40	3.01	28.42	2.04
Aug	74.95	2.57	17.89	3.52	27.07	2.62

Table 5: Ablations(%) of the full approach Ins.

5.4 Human Evaluation

Taking dialogue summarization as an example, we did human evaluation to further prove the improvement on sensitivity by sampling 200 pairs of generations for each offline approach and asked three proficient English speakers to label each case out of

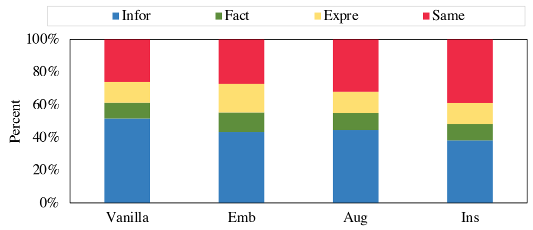
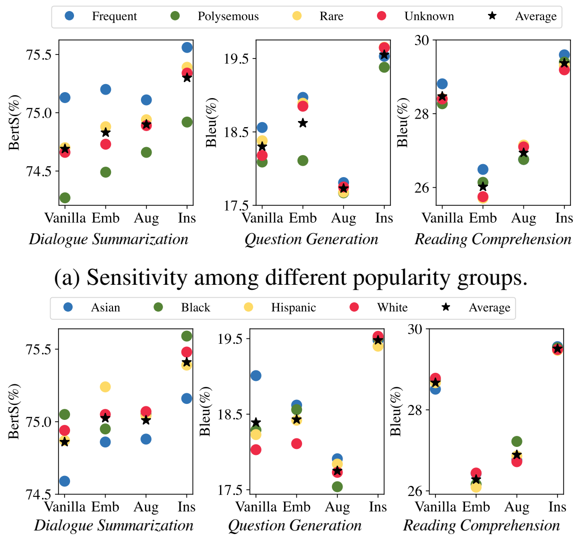


Figure 2: Human evaluation for difference types.



(a) Sensitivity among different popularity groups.

(b) Sensitivity among different racial groups.

Figure 3: Sensitivity among names within different groups. The scores are the higher the better and more centralized dots for each approach represent better insensitivity among groups.

4 choices by selecting the primary one that makes the generations distinct: **Infor**mation difference means both outputs contain different information or keywords. **Fact**ual difference refers to different matchings between speakers and events. **Expre**ssion difference is outputs having minor differences, such as capitalization and different orders of juxtaposed names. **Same** represents the identical outputs. The results are in Fig. 2 with 0.64 Kappa score, indicating substantial agreement. We can see that content distinction is the primary difference type. Ins generates less distinct contents and more identical results, outperforming the baselines.

5.5 Sensitivity among Name Groups

We collect specific groups of names in terms of popularity and race and show differences in the quality performances on test sets constructed with corresponding names. The sensitivity among different groups for each method are reflected by the scattering of dots vertically in Fig. 3.

Name groups by popularity and usage: We define 4 groups. **Frequent** including words frequently and solely used as human names is mentioned before. **Polysemous** represents words frequently used but not specialized for human names, such as June and Florida. **Rare** is names with low occurrence times like Paderau. **Unknown** names are similar to random strings from a model's perspective since they haven't been exposed to the model. The last three groups are collected by counting occurrences of all-possible names in the pre-training corpus of BART. We select 200 names for each group (More details are in Appendix B).

According to Fig. 3a, we can see that models usually perform poorly on Polysemous, even worse than Rare and Unknown. The daily meanings dominate the representation of this word and confuse the model. Frequent generally outperforms other groups. We conclude that words frequently and uniquely used as names that result in more specialized embeddings in pre-trained models and perform better. Moreover, comparing the sensitivity among different approaches, Ins outperforms the baselines in most cases except Aug. It achieves more centralized dots due to the performance reduction on the dominant groups or even all groups, showing that models tend to overfit with augmented data without our losses. To recap, Ins results in consistent improvements over Vanilla among different tasks compared with other baselines.

Name groups by races: Names from different races are from Tzioumis (2018) by assigning each name to a race with the highest probability. 4 major groups[4] are gathered, including Non-Hispanic **White**, **Hispanic** or Latino, Non-Hispanic **Black** or African American, and Non-Hispanic **Asian** or Native Hawaiian or Other Pacific Islander. To avoid the influence of the various number of names, we select the most frequent 50 names in each group and show the results in Fig. 3b. All of the approaches show discrimination against Asian in dialogue summarization. Emb, Aug and Ins improve the insensitivity among different races compared with Vanilla, and Ins is better with the guarantee on quality. We consider to introduce special designs on demographic features in the future.

[4] Other groups are empty after this assigning operation with Tzioumis (2018)'s name list.

Approach	R2				BertScore			
	-	S↓	R↓	D↓	-	S↓	R↓	D↓
In-distribution Names								
Vanilla	27.29	25.53	11.05	4.42	74.64	9.65	5.19	2.05
Emb	27.41	24.20	10.87	4.33	74.90	9.49	5.29	2.09
Aug	27.51	22.24	9.89	3.96	74.83	8.50	4.67	1.85
Ins⋆	**28.70**	**<u>16.54</u>**	**7.19**	**2.92**	**75.44**	**<u>6.11</u>**	**3.18**	**1.28**
All-possible Names								
Vanilla	27.32	23.77	11.07	4.45	74.81	9.61	5.15	2.04
Emb	27.26	24.98	10.68	4.25	75.30	9.57	5.16	2.02
Aug	27.36	22.73	10.04	4.03	74.86	8.56	4.69	1.87
Ins⋆	**28.38**	**<u>18.65</u>**	**8.12**	**3.29**	**75.35**	**<u>6.89</u>**	**3.75**	**1.50**

Table 6: Dialogue summarization results(%) of offline approaches for sensitivity on an individual speaker.

5.6 Sensitivity on an Individual Speaker

We can also only change the name of a single speaker each time to analyze fine-grained sensitivity. The results of offline approaches for dialogue summarization are shown in Table 6 (see more in Appendix D). The sensitivity scores are lower than the ones in Table 3. It seems that the sensitivity of models is proportional to the amount of changes in test samples, i.e., whether changing all speaker names (change-all-name) or only one speaker name (change-one-name). However, it's not always true and changing one name can be more sensitive than changing all names. Taking the results from Ins as an example, around 52.01% samples have speakers whose change-one-name D-BertS is higher than the corresponding changel-all-name one. Over 34.80% of the change-one-name D-BertS averaged by speakers from the same dialogue is also higher than the change-all-name D-BertS. We further show the trends between speaker features and their sensitivity scores in Fig. 4. Names are more sensitive and thus crucial for speakers at the start of a dialogue or with more utterances, deserving attention for further improvements.

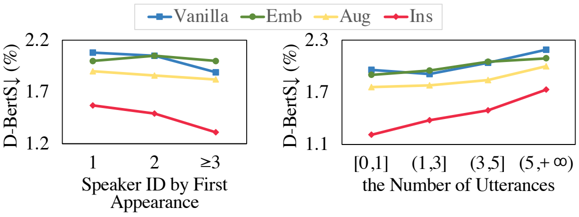


Figure 4: Change-one-name sensitivities on different speaker features for dialogue summarization.

6 Related Work

Entity/Name Bias in Narrative Texts: Previous work on entity biases shows that pre-trained language models are sensitive to changes in narrative text. Some works (Zhang et al., 2018, 2017; Wang et al., 2022b) for relation extraction mask entities in the context to prohibit learning spurious features between entities and relations. Yan et al. (2022) analyzes the robustness of models by entity renaming on reading comprehension. They all consider different kinds of entities, such as person and organization. However, the entities have the potential to be grounded in real life (Smith and Williams, 2021), and the background knowledge of these entities may be necessary for understanding. Besides, the context and the entities cannot always be well-separated, especially persons Yan et al. (2022). Thus, masking and switching operations are not always suitable for these entities. In our work, we focus on speakers that are not grounded.

Names that are not grounded have also been studied. Information such as age, gender and race can be reflected by a given name to some extent (Girma, 2020), while models learned with statistical features may make wrong predictions about specific persons or bring unexpected stereotypes (Bertrand and Mullainathan, 2004). Romanov et al. (2019) takes occupation classification as an example and discourages the model to predict an individual's occupation depending on his/her name. Wang et al. (2022a) presents that machine translation models perform poorly on female names when translating into languages with grammatical gender and also have sentiment bias caused by names with sentiment-ambiguous words. Samples in all these works only have a single name each, while multiple speaker names are entangled in a single dialogue.

Fairness of Dialogue Models: Safety and fairness issues on generations from dialogue models are crucial for implementation in practice. Harmful differences in responses caused by different demographic personas are observed in well-known dialogue systems (Sheng et al., 2021; Dinan et al., 2020), including offensiveness, gender bias, race discrimination, etc. These unfairness phenomena also exist in dialogue systems without considering persons (Liu et al., 2020), reflected by the politeness, sentiment, diversity and other aspects of a response. Recent work from (Smith and Williams, 2021) shows dialogue models treat their conversation partner differently for different speaker names. Instead of analyzing differences in open-ended dialogue systems, we target on text generation tasks given dialogues and show that sensitivity/unfairness also exists among speakers.

7 Conclusion

This paper focuses on the speaker name sensitivity in the text generation from dialogues. We provide a classification for previous approaches, and propose the insensitivity losses to reduce the sensitivity while achieving favorable generation quality. Fair comparisons and comprehensive analysis are done among different approaches for evaluating the sensitivity quantitatively. More approaches targeting dialogue sensitivity issues are expected.

Limitations

Our work has the following limitations:

First, we cannot generalize our conclusions to other languages that are dramatically different from English or more complicated multi-lingual scenarios without further experiments.

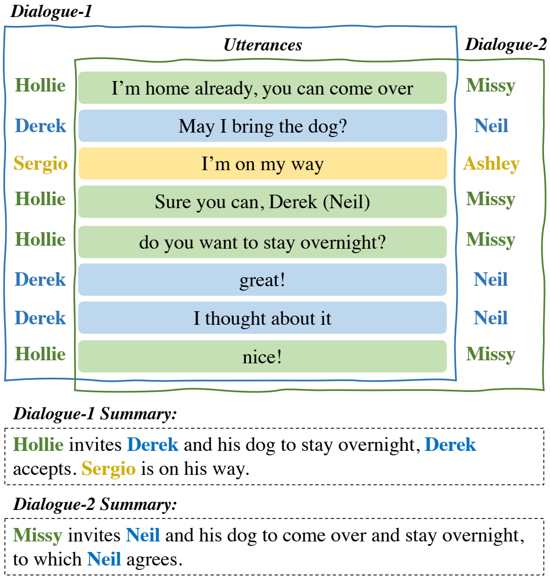
Second, we didn't consider any special designs on demographic features of names in our proposed approach. As shown in Sec. 5.5, discrimination does exist among different groups. Although Ins outperforms other baselines overall, there is still room to improve insensitivity among different groups for tasks with longer outputs containing multiple speaker names. We hypothesize that demographic features of names can be added through a more dedicated data augmentation strategy.

Third, our experimentation was restricted to the BART model in this paper. The reason is that among all the models that can be fine-tuned with our limited resources, including T5 and GPT-2, BART is still the best and the most popular, therefore we pick BART as the target of this study. Our intention is to devote the limited paper space to a more in-depth analysis of the problem using a range of tasks. Besides, it should be noticed that the speaker name sensitivity is still an issue with recent large pre-trained models, as shown in the example of dialogue summarization with outputs from ChatGPT in Fig. 5. The two summaries are expected to be the same, modulo speaker names. However, the third speaker (Sergio/Ashley) is not even mentioned in Summary-2.

We will try to address these limitations in the future.

Ethics Statement

All of the name lists we adopted in this paper are borrowed from public websites (https://www.ssa.gov) and previous publications (Tzioumis, 2018; Khalifa et al., 2021). We considered only binary genders and four different racial groups, which are clearly incomplete for depicting all humans. Our work is mainly at drawing researchers' attention to the unfairness caused by speaker names in text generation tasks given dialogues. These demographic features are selected to shed light on this potential issue and our method is not restricted to any specific demographic groups.



Dialogue-1	Utterances	Dialogue-2
Hollie	I'm home already, you can come over	Missy
Derek	May I bring the dog?	Neil
Sergio	I'm on my way	Ashley
Hollie	Sure you can, Derek (Neil)	Missy
Hollie	do you want to stay overnight?	Missy
Derek	great!	Neil
Derek	I thought about it	Neil
Hollie	nice!	Missy

Dialogue-1 Summary:

Hollie invites Derek and his dog to stay overnight, Derek accepts. Sergio is on his way.

Dialogue-2 Summary:

Missy invites Neil and his dog to come over and stay overnight, to which Neil agrees.

Figure 5: An example of dialogue summarization with outputs from ChatGPT.

Acknowledgments

This work was generously supported by the CMB Credit Card Center & SJTU joint research grant, and Meituan-SJTU joint research grant.

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188.

Hewan Girma. 2020. Black names, immigrant names: Navigating race and ethnicity through personal names. *Journal of Black Studies*, 51(1):16–36.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker turn modeling for dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2150–2157. Association for Computational Linguistics.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen Mckeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022.

Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. 2019. The eighth dialog system technology challenge. *arXiv preprint arXiv:1911.06394*.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. *arXiv preprint arXiv:1904.05233*.

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing persona biases in dialogue systems. *arXiv preprint arXiv:2104.08728*.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. pages 6850–6861.

Eric Michael Smith and Adina Williams. 2021. Hi, my name is martha: Using names to measure and mitigate bias in generative dialogue models. *arXiv preprint arXiv:2109.03300*.

Konstantinos Tzioumis. 2018. Demographic aspects of first names. *Scientific data*, 5(1):1–9.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022a. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2576–2590.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022b. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *Proceedings of the 8th International Conference on Learning Representations*.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

A Illustration for Insensitivity Losses

Fig. 6 depicts the positions of the cross attentions and the final decoder hidden states in the encoder-decoder Transformer model for a better understanding of our two insensitivity losses.

B Name Groups

To collect polysemous, rare and unknown names, we counted the number of occurrences of all-possible names in the pre-training corpus, Wikipedia[5] and BookCorpus[6]. We denote the frequency of a name as f_{exact} and f_{ner} representing doing exact string match or named entity recognition when counting name occurrences respectively. Rare contains names shown at least once and with the lowest f_{exact} not equaling 0. Unknown includes names with f_{exact} equaling 0. According to our observations, we find that names with a larger f_{exact} are likely to be polysemy and are not uniquely used as personal names. So, we design a metric to recognize such names as follows:

$$u = \frac{rank(f_{exact}) - rank(f_{ner})}{rank(f_{exact}) + rank(f_{ner})} \tag{9}$$

$rank(\cdot)$ means that the ranking of a name among the whole name list based on its frequency in descending order [7]. A higher u shows a higher level of uniqueness of a word as a name. The names with the lowest u scores are selected as Polysemous in Sec. 5.5.

Examples of names in different name groups are listed as follows:

- **Frequent**: Alexis, Philip, Matthew, Frank, Tyler, Roy, Catherine, Joan, Amanda, Henry
- **Polysemous**: July, Sea, March, Paris, Treasure, Oxford, Romania, Ice, Jersey, Navy
- **Rare**: Makinzy, Diyanna, Javione, Zamire, Harkeem, Jerralyn, Crissi, Monque, Ajahar, Dijion
- **Unknown**: Jaliyiah, Cardelia, Ravindr, Josephanthony, Tyjohn, Tnaya, Jyren, Kashaunda, Jaykob, Latonnia
- **White**: Kim, Georgia, Joseph, Mark, Martin, James, William, Barbara, Richard, Victoria
- **Hispanic**: Sofia, Daisy, Luis, Manuel, Dora, Emilia, Minerva, Antonio, Oscar, Francisco
- **Black**: Kenya, Ebony, Anderson, Kelvin, Dexter, Cleveland, Percy, Mamie, Jarvis, Essie
- **Asian**: Kong, Muhammad, Gang, Mai, Chi, Krishna, Can, Wan, Wang, Ferdinand

[5] https://huggingface.co/datasets/wikipedia

[6] https://huggingface.co/datasets/bookcorpus

[7] Doing named entity recognition on the whole pre-training corpus is too time-consuming. Therefore, we randomly sample 1% of the data for counting the f_{ner} and use the name rankings in Eq. 9 to get the uniqueness score

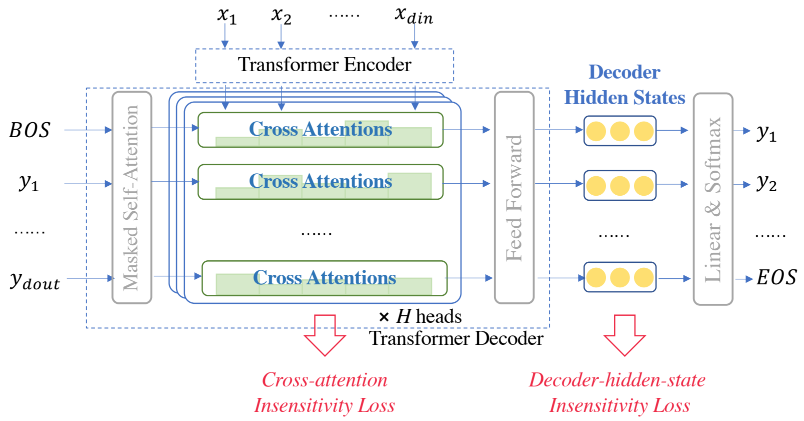


Figure 6: An illustration of insensitive losses. BOS and EOS are special tokens standing for the start and the end of the output.

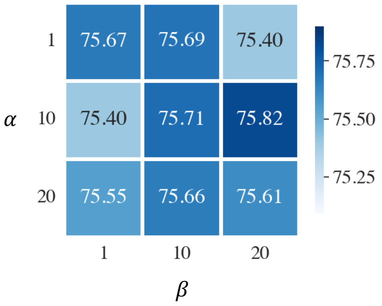


Figure 7: BertScore(%) on the vanilla test set with different hyper-parameters.

C Hyper-parameter Search

We empirically searched the hyper-parameters α and β in {1, 10, 20} respectively with 9 combinations for Ins. Due to the limited computation resources and the large search space, we trained the model with different combinations for a single time, selected the best 3 combinations and repeated experiments with different random seeds to determine the final choice of α and β according to the performance on D_{va}. Finally, we set (α, β) as (1, 10), (1, 10), (1,1) for dialogue summarization, question generation and reading comprehension respectively. We directly borrow these settings for FreIns.

In Fig. 7, we show the performances of Ins under different combinations for dialogue summarization on the vanilla test set with a single run. We can see that all of the results outperform the baselines in Table 2 and the standard deviation of BertScore among different combinations is only 0.14%, showing the stable improvements of Ins over the baselines.

D Additional Results of Sensitivity on an Individual Speaker

Results for sensitivity on an individual speaker on all of the three tasks are in Table 7 and Table 8. Both tables lead to the same observations and conclusions as discussed in Sec 5.1 and Sec 5.2, where Ins and FreIns perform best among offline and online approaches correspondingly.

E Case study

We show cases for different tasks in this section.

The case for dialogue summarization is in Fig. 8. Vanilla extracts different information for two sets of names: “She will bring eggs” and “Ethie is off on Friday”. It also uses different expressions: “will come to ... for Easter” and “invited ... for Easter”. Besides, “Louise” is only mentioned in the second summary. Emb has the information difference and the expression difference. Meanwhile, it outputs incorrect content in the second summary, where “chocolat ones” is used for describing “eggs” in the input dialogue. Aug outputs more information for the first set of names. Ins treats the two sets of names equally with the same generations modulo the speaker names.

In the case of question generation in Fig. 9, all baselines generate “who gives Jernee suggestions?” for the second set of names, which is an inaccurate question with multiple candidate answers. Emb also generates a “Who” with the capitalized first letter, which is also different from the other one with lowercase “who” if we compare them strictly. Ins generates identical and accurate questions for the same dialogue with different speaker names.

For reading comprehension in Fig. 10, both

Approach	R2				BertScore			
	-	S↓	R↓	D↓	-	S↓	R↓	D↓
In-distribution Names								
Vanilla	27.29	25.53	11.05	4.42	74.64	9.65	5.19	2.05
Emb	27.41	24.20	10.87	4.33	74.90	9.49	5.29	2.09
Aug	27.51	22.24	9.89	3.96	74.83	8.50	4.67	1.85
Ins⋆	**28.70**	**16.64**	**7.19**	**2.92**	**75.44**	**6.11**	**3.18**	**1.28**
All-possible Names								
Vanilla	27.32	25.77	11.07	4.45	74.81	9.61	5.15	2.04
Emb	27.26	24,98	10.68	4.25	74.80	9.57	5.16	2.02
Aug	27.36	22.73	10.04	4.03	74.86	8.56	4.69	1.87
Ins⋆	**28.38**	**18.65**	**8.12**	**3.29**	**75.35**	**6.89**	**3.75**	**1.50**

(a) Dialogue Summarization

Approach	BLEU				RL			
	-	S↓	R↓	D↓	-	S↓	R↓	D↓
In-distribution Names								
Vanilla	17.93	18.76	6.08	2.58	56.85	8.17	7.55	3.12
Emb	18.34	22.22	7.63	3.26	56.84	10.07	9.62	3.98
Aug	18.06	14.82	4.39	1.90	56.12	6.91	6.38	2.69
Ins⋆	**19.45**	**9.66**	**2.75**	**1.18**	**57.31**	**4.50**	**4.27**	**1.81**
All-possible Names								
Vanilla	17.91	17.73	5.75	2.46	56.67	7.76	7.05	2.95
Emb	18.67	20.80	7.08	3.06	56.86	9.47	8.89	3.73
Aug	17.97	13.04	3.62	1.57	56.12	6.06	6.50	2.25
Ins⋆	**19.60**	**8.11**	**2.22**	**0.97**	**57.51**	**3.77**	**3.42**	**1.47**

(b) Question Generation

Approach	BLEU				RL			
	-	S↓	R↓	D↓	-	S↓	R↓	D↓
In-distribution Names								
Vanilla	27.96	54.08	3.85	1.67	73.91	4.49	5.50	2.37
Emb	25.52	56.61	4.28	1.85	70.20	5.32	6.37	2.75
Aug	26.54	54.76	3.69	1.60	72.53	4.57	5.87	2.55
Ins⋆	**29.03**	**52.03**	**2.48**	**1.08**	**74.81**	**5.65**	**4.41**	**1.91**
All-possible Names								
Vanilla	27.82	53.48	2.81	1.22	73.97	3.28	4.07	1.77
Emb	25.14	56.08	3.04	1.32	70.51	4.31	4.89	2.12
Aug	26.64	53.71	2.92	1.27	72.68	3.61	4.61	2.00
Ins⋆	**29.40**	**51.20**	**1.93**	**0.83**	**74.94**	**2.41**	**3.13**	**1.36**

(c) Reading Comprehension

Table 7: Performances(%) of offline approaches for sensitivity on an individual speaker.

Approach	R2				BertScore			
	-	S↓	R↓	D↓	-	S↓	R↓	D↓
Fre	28.40	20.28	9.25	3.73	75.10	7.85	4.29	1.72
FreAug	27.91	20.11	9.02	3.64	74.97	7.78	4.24	1.70
FreIns⋆	**28.58**	**13.29**	**5.99**	**2.46**	**75.42**	**4.91**	**2.68**	**1.09**

(a) Dialogue Summarization

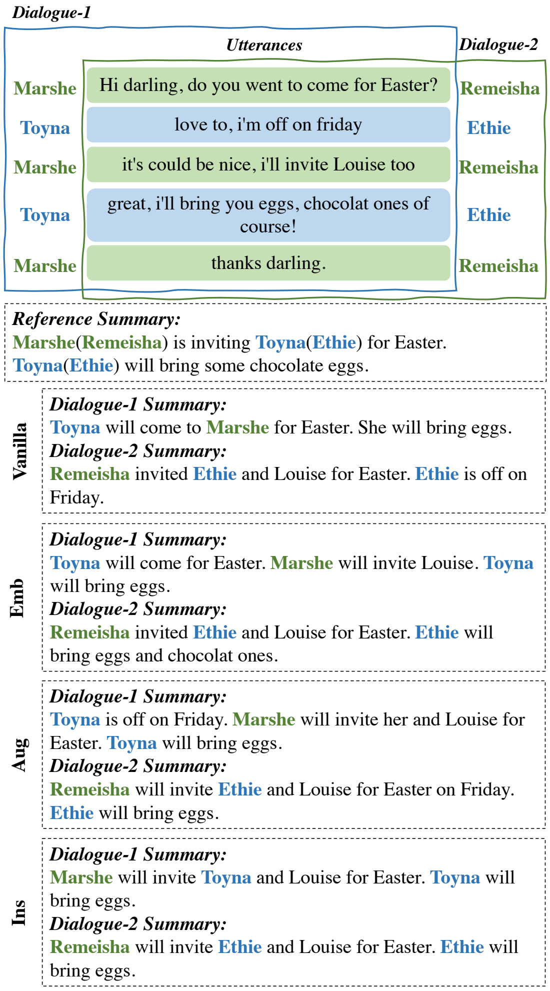
Approach	BLEU				RL			
	-	S↓	R↓	D↓	-	S↓	R↓	D↓
Fre	18.90	10.59	2.97	1.29	57.01	4.76	4.09	1.74
FreAug	18.60	8.62	2.54	1.10	57.13	3.81	3.46	1.48
FreIns⋆	**19.29**	**5.48**	**1.76**	**0.77**	**56.91**	**2.39**	**2.18**	**0.94**

(b) Question Generation

Approach	BLEU				RL			
	-	S↓	R↓	D↓	-	S↓	R↓	D↓
Fre	27.15	53.86	2.07	0.89	73.89	2.67	3.22	1.39
FreAug	27.82	**52.03**	1.83	0.80	74.32	2.33	3.08	1.33
FreIns⋆	**28.57**	52.41	**1.46**	**0.64**	**74.89**	**1.70**	**2.36**	**1.02**

(c) Reading Comprehension

Table 8: Performances(%) of online approaches for sensitivity on an individual speaker.



Dialogue-1	Utterances	Dialogue-2
Marshe	Hi darling, do you went to come for Easter?	Remeisha
Toyna	love to, i'm off on friday	Ethie
Marshe	it's could be nice, i'll invite Louise too	Remeisha
Toyna	great, i'll bring you eggs, chocolat ones of course!	Ethie
Marshe	thanks darling.	Remeisha

Reference Summary:
Marshe(Remeisha) is inviting Toyna(Ethie) for Easter.
Toyna(Ethie) will bring some chocolate eggs.

Vanilla

Dialogue-1 Summary:
Toyna will come to Marshe for Easter. She will bring eggs.
Dialogue-2 Summary:
Remeisha invited Ethie and Louise for Easter. Ethie is off on Friday.

Emb

Dialogue-1 Summary:
Toyna will come for Easter. Marshe will invite Louise. Toyna will bring eggs.
Dialogue-2 Summary:
Remeisha invited Ethie and Louise for Easter. Ethie will bring eggs and chocolat ones.

Aug

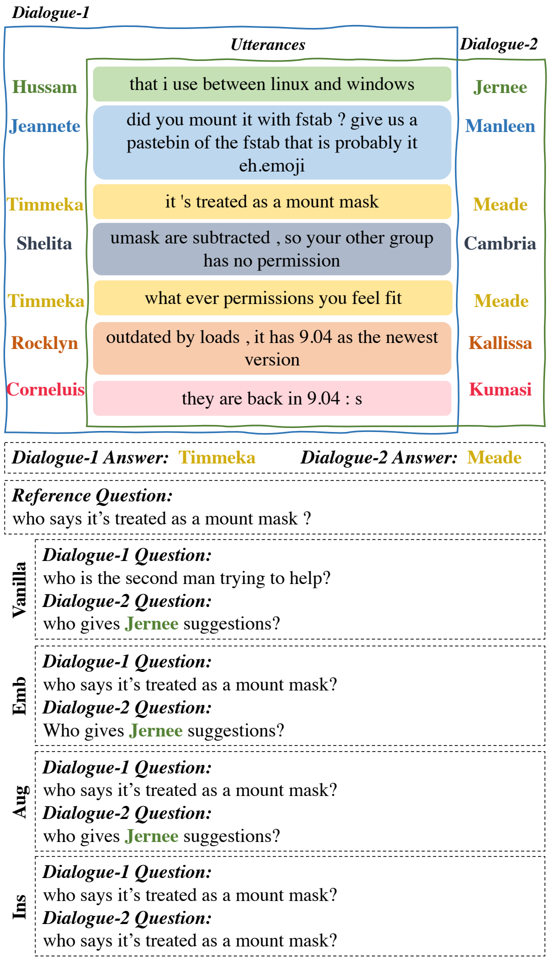
Dialogue-1 Summary:
Toyna is off on Friday. Marshe will invite her and Louise for Easter. Toyna will bring eggs.
Dialogue-2 Summary:
Remeisha will invite Ethie and Louise for Easter on Friday. Ethie will bring eggs.

Ins

Dialogue-1 Summary:
Marshe will invite Toyna and Louise for Easter. Toyna will bring eggs.
Dialogue-2 Summary:
Remeisha will invite Ethie and Louise for Easter. Ethie will bring eggs.

Figure 8: Case study for dialogue summarization.

Vanilla and Emb generate quite different answers for two sets of names. Aug generates consistent but wrong answers considering the one-to-one mapping of speaker names. Ins outputs identical correct and complete answers, outperforming the baselines.



Dialogue-1

Dialogue-1	*Utterances*	*Dialogue-2*
Hussam	that i use between linux and windows	Jernee
Jeannete	did you mount it with fstab ? give us a pastebin of the fstab that is probably it eh.emoji	Manleen
Timmeka	it 's treated as a mount mask	Meade
Shelita	umask are subtracted , so your other group has no permission	Cambria
Timmeka	what ever permissions you feel fit	Meade
Rocklyn	outdated by loads , it has 9.04 as the newest version	Kallissa
Corneluis	they are back in 9.04 : s	Kumasi

Dialogue-1 Answer: Timmeka ***Dialogue-2 Answer:*** Meade

Reference Question:
who says it's treated as a mount mask ?

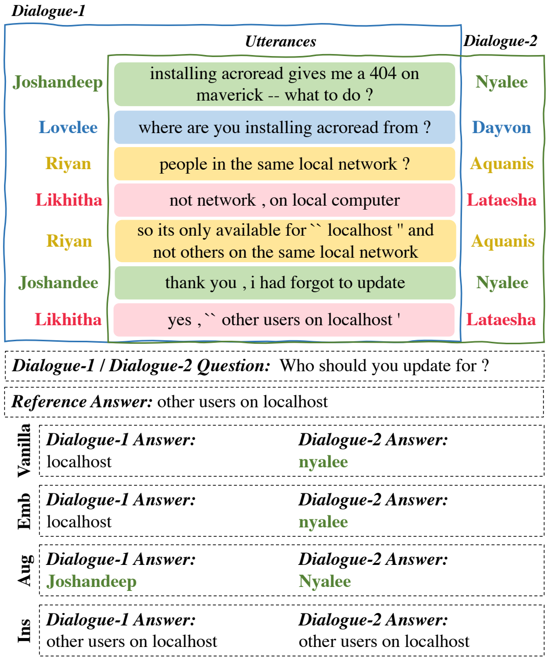
Vanilla
Dialogue-1 Question:
who is the second man trying to help?
Dialogue-2 Question:
who gives **Jernee** suggestions?

Emb
Dialogue-1 Question:
who says it's treated as a mount mask?
Dialogue-2 Question:
Who gives **Jernee** suggestions?

Aug
Dialogue-1 Question:
who says it's treated as a mount mask?
Dialogue-2 Question:
who gives **Jernee** suggestions?

Ins
Dialogue-1 Question:
who says it's treated as a mount mask?
Dialogue-2 Question:
who says it's treated as a mount mask?

Figure 9: Case study for question generation.



Dialogue-1

Dialogue-1	*Utterances*	*Dialogue-2*
Joshandeep	installing acroread gives me a 404 on maverick -- what to do ?	Nyalee
Lovelee	where are you installing acroread from ?	Dayvon
Riyan	people in the same local network ?	Aquanis
Likhitha	not network , on local computer	Lataesha
Riyan	so its only available for `` localhost '' and not others on the same local network	Aquanis
Joshandee	thank you , i had forgot to update	Nyalee
Likhitha	yes , `` other users on localhost '	Lataesha

Dialogue-1 / Dialogue-2 Question: Who should you update for ?

Reference Answer: other users on localhost

	Dialogue-1 Answer:	***Dialogue-2 Answer:***
Vanilla	localhost	**nyalee**
Emb	localhost	**nyalee**
Aug	**Joshandeep**	**Nyalee**
Ins	other users on localhost	other users on localhost

Figure 10: Case study for reading comprehension.